

Multi-Armed Bandit Problems with Delayed Feedback

Sudipto Guha*

Kamesh Munagala[†]

Martin Pál[‡]

Abstract

In this paper we initiate the study of optimization of bandit type problems in scenarios where the feedback of a play is not immediately known. This arises naturally in allocation problems which have been studied extensively in the literature, albeit in the absence of delays in the feedback. We study this problem in the Bayesian setting. In presence of delays, no solution with provable guarantees is known to exist with sub-exponential running time.

We show that bandit problems with delayed feedback that arise in allocation settings can be forced to have significant structure that gives us the ability to reason about this policy. We show a $O(1)$ approximation for a significantly general class of priors. The structural insights we develop are of key interest and carry over to the setting where the feedback of an action is available instantaneously. In particular, we show a simple 2-approximation for the finite horizon Bayesian bandit problem, improving and generalizing prior work.

*Research supported in part by an NSF Award CCF-0644119 and IIS-0713267. Part of this work was done while the author was visiting Google Research. Department of Computer and Information Sciences, University of Pennsylvania, Philadelphia, PA. Email: sudipto@cis.upenn.edu

[†]Supported by an Alfred P. Sloan Research Fellowship, and by NSF via CAREER award CCF-0745761 and grant CCF-1008065. Department of Computer Science, Duke University, Durham, NC 27708-0129. Email: kamesh@cs.duke.edu

[‡]Google, 76 9th Ave, New York, NY. Email: mpal@google.com

1 Introduction

In this paper, we consider the problem of iterated allocation of resources, when the effectiveness of a resource is uncertain, and is learnt after some delay. Allocation of resources under uncertainty is a central problem in a variety of disciplines, notably in learning [7] and stochastic control [6]. In these problems, we are asked to make a series of allocation decisions, based on past outcomes. Since the seminal contributions of Wald [31] and Robbins [26], a vast literature, including both optimal and near optimal solutions, has been developed, see references in [4, 30, 6, 7]. Of particular interest is the celebrated Multi-Armed Bandit (MAB) problem, where an agent decides on allocating resources between competing actions (arms) with uncertain rewards and can only take one action at a time (play the arm). The agent collects the reward and the state of the played arm is updated.

However, an overwhelming majority of the literature focuses on scenarios that assume *instantaneous* (or negligible in comparison to the horizon) revelation of the outcome of each allocation. In an early work in mid 1960s, Anderson [2] introduced the notion of delayed observations. Since then, though there have been additional results [29, 9], a theoretical guarantee on adaptive decision making under delayed observations has been elusive, and the computational difficulty in obtaining such has been commented upon in [12, 3, 28, 13]. Recently, this issue of delays has been thrust to the fore due to the increasing application of iterative allocation problems in online advertising.

1.1 Problem Statement and Motivating Examples

We now concretely define the Bayesian multi-armed bandit problem with delayed feedback. There is a bandit with n independent arms. When arm i is played, the reward is drawn *i.i.d.* from distribution D_i , which is unknown. However, a prior \mathcal{D}_i is specified over possible D_i . When arm i is played, the feedback about its reward outcome is learned only after δ_i steps. The set of observed outcomes so far resolves the prior to a posterior distribution according to Bayes' rule. We also assume each arm has budget B_i on the maximum reward it can accrue; further plays do not accrue reward. In this setting, the goal is to design a **decision policy** for allocating the plays to the arms. A decision policy is a mapping from the current *state*, determined by the posterior distributions of each arm; the plays for which feedback is outstanding; and the remaining budgets of each arm to an action of which arm to play. There is a horizon of T steps, and our goal is to design a polynomial time algorithm that outputs a decision policy maximizing the expected revenue.

As concrete instantiations of this framework and a running example of the space of problems we address, consider the following three examples:

1. Online Determination of Web Content: This is described by Agarwal et al in [1]. The goal is to present different websites/snippets to an user and induce the user to visit these pages. For each user (type) j there is prior information about the propensity of the user to be interested in content i . This propensity is described by a distribution D_{ij} drawn from a prior distribution \mathcal{D}_{ij} . The priors are constructed from historical estimation. As the pages are displayed, we resolve this prior; however, the characterization of the goodness a visit corresponds to multiple signals which is only known after a delay δ_{ij} . The delay can also arise from batched updates and systems issues. Note that there are a number of “budget constraints” where each page can be displayed T_j times. For one user, this maps to the problem described above. The authors of [1] present a number of different heuristics based on greedy policies.

2. Unmanned Aerial Vehicles: Bandit problems are used widely in stochastic control, where often actions are not immediate in their feedback. For example the difficulty in routing multiple UAVs among an uncertain terrain is the remoteness of the controller [24], which induces delays.

3. Budgeted Allocation Problem: In this problem, there are n bidders (advertisers competing for a collection of m impressions (advertisement slots)). Assume that the impressions arrive one at a time, and correspond to the "semi-online" framework where the number of impressions of type j is T_j and known beforehand. Each bidder i has a budget B_i on the total amount she is willing to pay, and bids b_{ij} for impression of type j . When advertiser i is allocated an impression of type j , she is charged her bid b_{ij} only if an acquisition occurs. This is referred to as the cost-per-action (CVR) model. The search engine allocates the impressions based on its estimate of the click or conversion probability (CVR) p_{ij} of this favorable event, in order to maximize its expected revenue (where if a favorable event occurs for pair (i, j) , and i 's budget is not exhausted, the contribution to revenue is b_{ij}). The classical budgeted allocation problem [23, 15, 8, 11] assumes that (i) the p_{ij} are known; and (ii) feedback about the click or conversions is instantaneous. Stochastic relaxations of the assumption (i) is already considered in [25], who also provide simulations of greedy algorithms and consider budgets. But in case of conversions (and in case of more refined measurements of clicks as well) there is a natural delay in the feedback, which previous literature largely ignores.

1.2 Our Results and Techniques

We present a constant factor approximation to the Bayesian MAB problem with delayed feedback. Before presenting our results, it is instructive to study closely related work. The Bayesian MAB problem has been considered in the model where the feedback is instantaneous, see [27, 22, 18, 17], and constant factor approximations are known in this setting. In the absence of delays, all approximation algorithms (as well as the exact Gittins index algorithm in the discounted reward setting [16, 30]) use the key observation that the posterior distribution of an arm does not change if not played. Using this observation, it was shown in [18] that there is an approximately optimal policy that sequentially plays the arms. In presence of delays, the state of an arm now needs to capture the plays with outstanding feedback, and this not only causes the state to change even with no plays, but also requires exponential space to specify the state even for a single arm. Finally, any policy has to now interleave plays of the different arms, and hence, existing techniques in [18, 17, 30] do not apply in any straightforward fashion. In a sense, the challenge in the delayed setting is not just to balance explorations along with the exploitation, *but also to decide a **schedule** for the possible actions both for a single arm (due to delays) as well as across the arms.*

At a high level, our algorithm starts with the standard LP relaxation to the problem [6, 18]. This relaxation provides a collection of *single-arm* policies whose execution is confined to one bandit arm; the LP enforces constraints on the total expected plays these policies can make. Such a relaxation has been widely used in the absence of delays, since it leads to efficient **decomposable decision policies** (called index policies) that treats every arm independently of other arms. Note that the joint state space over all the arms is exponential in size. As alluded to above, in the presence of delays, it is not even clear that such decomposable policies are efficiently computable. The most important question that we ask for delayed feedback therefore is: *Do there exist near optimum policies that are decomposable and efficiently computable?* In this paper we develop new techniques to analyze single-arm policies and provide a poly-time algorithm that outputs a decomposable decision policy¹ which is an $O(1)$ approximation of the possibly entangled optimum policy.

In the rest of the paper we focus on the MAB case; we show in Appendix A how the results generalizes to multiple MABs or users (budgeted allocation case). This generalization is a consequence of the fact that the algorithm uses the decomposition property (which implies linearity).

We summarize our main results and technical contributions below:

¹These policies are very close to index policies; the policy of each arm is independent of other arms and hence efficiently computable; however, the policies of different arms cannot be reduced to a single *priority* value.

- In Section 3, we prove a structural result, *Truncation Theorem*. Theorem 1 shows that for arbitrary state spaces, even in the presence of budgets, delays and other constraints which preserves paths, a partially executed single-arm policy has reward proportional to the original policy. This uses a stopping time type argument and uses the connection between state spaces corresponding to priors and their Martingale properties. This idea leads to shortening the horizon of single-arm policies, and is very general; it directly yields a 2-approximation for the finite horizon MAB problem with instantaneous feedback. In the absence of budgets, the best known result for this problem was a 12 approximation [18, 17]; their approach is infeasible in the presence of budgets. Our result also shows a factor 2 approximation for the "irrevocable bandit" problem², improving the factor 8 result of Farias and Madan [14]. We also show that factor 2 is the best possible bound against LP relaxations over single arm policies. This result has also been used in subsequent results for MABs with nonlinear objective functions [19].
- In Section 4, using the Truncation theorem, we show a $O(1)$ approximation for the MAB problem with delayed feedback when $\delta_i = o(T/\log T)$.³ Applying the Truncation theorem is not straightforward - a policy could make very few plays initially waiting for feedback after every play, and load all the plays at the very end when it is fairly certain the reward is large. To circumvent this, we develop an interesting compaction and simulation argument - we modify the policy to make more plays upfront, but withhold using the outcomes until the original policy used them. The resulting structure also implies that our policies are as efficient to compute as standard index policies (without delays) [16, 6]. Finally, the combination of the single-arm policies into a final feasible policy requires a novel priority based scheme.

The chief technical highlight of our work is a new way of accounting for the reward of a single-arm policy using the martingale property of priors that Bayes' rule entails. Traditionally, the Markov Decision Process (MDP) formulation resulted in the reward being accounted non-uniformly: given a posterior distribution, a play yields reward which is the expected value of the posterior. We use the fact that the plays draw *i.i.d.* rewards from a fixed underlying distribution (whose current belief is encoded in the prior), and hence the expected reward is the same for each play conditioned on this unknown distribution. We use this accounting for both the truncation and compaction steps, leading to simplicity of policies, clarity of analysis, and improved approximation bounds.

1.3 Other Related Work

There is an extensive literature on the MAB problem in the prior-free setting; see [21, 4, 7]), and policies with additive *regret* guarantees are known. Regret is the difference between the expected reward of the policy and the reward of an omniscient policy which knows all the distributions. However, these results both require the reward rate to be large and large time horizon T compared the number of arms. In the application scenarios mentioned above, it will typically be the case that the number of arms is very large and comparable to the optimization horizon and the reward rates are low. This motivates the need for a purely multiplicative guarantee instead of additive guarantees. Moreover the analysis of these policies require the plays of the arm with maximum estimated reward to be continuous, which is not true in presence of delays (or budgets). If the delay of arm i satisfies $\delta_i = o(n^{1/3}T^{2/3})$ (where n is the number of arms) then standard explore-then-exploit (where we play an arm long enough to start receiving the outcomes) strategy gives sub-linear in T regret. However, constants in the regret term depends on scaling of the rewards.

²The algorithm is not allowed to revisit an arm.

³We can show that using stochastic domination of Gittins-type indices, we can increase δ_i to $O\left(\frac{T}{\log \log T}\right)$ and still preserve the $O(1)$ approximation of decomposable policies. However, the proof is very technical and omitted.

2 Preliminaries

There is a bandit with n independent arms. The arm i underlying reward distribution D_i , which is a random variable drawn from a prior distribution \mathcal{D}_i . These priors are specified as input. The maximum possible reward that can be extracted from this arm is B_i , and though additional plays can be made, they do not accrue additional reward. If an arm is played, the feedback about the reward outcome is available only after δ_i time steps. As observations are available the successive posteriors (which serve as priors for the next trial), are produced by the Bayes' Rule. There is a time horizon of T plays. A **decision policy** specifies which arm to play given the current state of each arm, which is captured by the remaining budget and time horizon, posterior distribution, and plays with outstanding feedback for that arm. Each decision policy has a unique expected reward value that is obtained on executing it (with different execution trajectories differing due to different realizations of the underlying $\{D_i\}$). **Our goal** is to design a poly-time algorithm, which outputs a policy that approximately maximizes the expected value derived over the horizon of T plays.

Input: Priors and Posterior Spaces. For each arm i , the input specifies the space \mathcal{S}_i of possible posterior distributions, which defines a natural DAG, the root of which corresponds to the initial prior \mathcal{D}_i . Every other state $u \in \mathcal{S}_i$ corresponds to a set of observations, and hence to the posterior X_{iu} obtained by applying Bayes rule to \mathcal{D}_i with the those observations. Playing the arm in state u yields a transition to state v with probability \mathbf{p}_{uv} , provided v can be obtained from u in one additional observation; the probability \mathbf{p}_{uv} is simply the probability of this observation conditioned on the posterior X_{iu} at u . The expected posterior mean at a state $u \in \mathcal{S}_i$, denoted by $r_u = \mathbf{E}[X_{iu}]$ satisfies the **martingale** property $r_u = \sum_v \mathbf{p}_{uv} r_v$. An example is in Appendix B.

We need the running time to be $\text{poly}(n, T, \sum_i |\mathcal{S}_i|)$, which is comparable to the running times (via per-arm dynamic programming) for computing the standard index policies [6, 16].

Budgets: The budget B_i for arm i can be folded into the description of the posterior space: If the observations leading to the current posterior already violate the budget, the reward of this state is set to 0. This transformation helps us ignore the budget in subsequent sections.

Single-arm Policies. Given an execution of the global policy \mathcal{P} , define its **projection** on arm i to be the policy \mathcal{P}_i defined by the actions induced on \mathcal{S}_i ; we term this a *single-arm* policy. Note that the global policy may take an action in arm i based on information regarding other arms (an entangled state) – that side information is lost in the projection.

Let $\{\mathcal{P}_s(i, T)\}$ describe all single-arm policies of arm i with feedback delay of δ_i steps and horizon T . Each $\mathcal{P}_s(i, T)$ is a (randomized) mapping from the current state to one of the following actions: (i) make a play; (ii) wait some number of steps (less or equal to T), so that when the result of a previous play is known, the policy changes state; (iii) wait a few steps and make a play (without extra information); or (iv) quit. The *state* of the system is captured by the current posterior $u \in \mathcal{S}_i$, the plays with outstanding feedback, and the remaining time horizon. Note that the state encodes plays with outstanding feedback, and this has size 2^{δ_i} , which is exponential in the input. Ameliorating this dependence is an aspect that we address in our algorithm design.

Definition 1. *Given a single-arm policy \mathcal{P}_i let $R(\mathcal{P}_i)$ to be the expected reward and $N(\mathcal{P}_i)$ as the expected number of plays, where the expectation is over the outcomes of the plays of \mathcal{P}_i .*

3 The Truncation Theorem

We now show that the time horizon of a single-arm policy on T steps can be reduced to βT for constant $\beta \leq 1$ while sacrificing a constant factor in the reward. We note that though the statement

seems simple, this theorem *only* applies to single-arm policies and is *not true* for the global policy executing over multiple arms. The proof of this theorem uses the martingale structure of the rewards, and proceeds via a stopping time argument.

Theorem 1. (*The Truncation Theorem*) For any arbitrary single arm policy \mathcal{P} which traces out a path over a horizon of T steps in a state space \mathcal{S} , then the identical policy \mathcal{P}' that makes at least a β fraction of the plays on any decision path, satisfies: (i) $R(\mathcal{P}') \geq \beta R(\mathcal{P})$ and (ii) $N(\mathcal{P}') \leq N(\mathcal{P})$.

Proof. Let the average reward in the initial state be μ , and the prior is characterized by a distribution $f(\mu)$. Consider the tree defined by the policy \mathcal{P} . Let $R(\mathcal{P}(\mu)), N(\mathcal{P}(\mu))$ denote the expected reward and the number of plays of the policy \mathcal{P} when the average reward is fixed at μ . We have $R(\mathcal{P}) = \int R(\mathcal{P}(\mu))f(\mu)d\mu$. **The critical part** of the proof is the next claim which basically reduces to the fact that the decision to play does not affect the outcome.

Claim 1. If we execute the policy $\mathcal{P}(\mu)$ from any node $u \in \mathcal{S}$ the expected reward $R(\mathcal{P}(\mu)) = \sum_v \mu y_\mu(v) \text{length}(v)$ where $\{y_\mu(v)\}$ is the distribution induced on the vertices of \mathcal{S} where $\mathcal{P}(\mu)$ stops executing. The $\text{length}(v)$ refers to the length of the path (in number of edges) from u to v .

Proof. We observe that if we play at a node u' , then we generate a reward μ . Now the probability of being at u' is the sum of the probability of the paths that pass through u' .

$$R(\mathcal{P}(\mu)) = \sum_{u'} \mu \Pr[\text{reaching } u'] = \sum_{u'} \sum_{v: u' \text{ is on a path to } v, u' \neq v} \mu y_\mu(v) = \sum_v \mu y_\mu(v) \text{length}(v)$$

The claim follows. \square

(Continuing Proof of Theorem 1) Truncating $\mathcal{P}(\mu)$ to $\mathcal{P}'(\mu)$ induces a many to one map over the paths at which the policy stops. Let $\{y_\mu(v)\}, \{y'_\mu(v')\}$ be the distribution induced on the vertices of \mathcal{S} where $\mathcal{P}(\mu), \mathcal{P}'(\mu)$ stop executing, respectively.

If the path to v got truncated to $g(v)$, then $y'_\mu(v') = \sum_{v: g(v)=v'} y_\mu(v)$. Further, from the statement of the Lemma, $\text{length}(g(v)) = \beta T \geq \beta \text{length}(v)$. Then using Claim 1, on $\mathcal{P}'(\mu)$,

$$\begin{aligned} R(\mathcal{P}'(\mu)) &= \sum_{v'} \mu y'_\mu(v') \text{length}(v') = \sum_{v'} \mu \sum_{v: g(v)=v'} y_\mu(v) \text{length}(v') = \sum_{v'} \mu \sum_{v: g(v)=v'} y_\mu(v) \text{length}(g(v)) \\ &\geq \sum_{v'} \mu \sum_{v: g(v)=v'} y_\mu(v) \beta \text{length}(v) = \beta \sum_v \mu y_\mu(v) \text{length}(v) = \beta R(\mathcal{P}(\mu)) \end{aligned}$$

The last part follows from applying Claim 1 on $R(\mathcal{P}(\mu))$. Note that the rewards of each path is *not* preserved up to a factor. But the Claim 1 introduces an accounting method where the reward is redistributed over the length. Note that it was important that the outcome of the play did not depend on the decision to play. Integrating over μ , the result of $R(\mathcal{P}')$ follows. The truncation can not increase the expected number of plays, this proves the theorem. \square

3.1 Finite Horizon MAB with Instantaneous Feedback

As a direct application, we consider the Finite Horizon Bayesian Problem in the delay free setting. To formulate an LP relaxation, we find the optimal single-arm policy for each arm, so that the resulting ensemble of policies have expected number of plays at most T . This yields:

$$LP1 = \max_{\mathcal{P}_s(i, T)} \left\{ \sum_i R(\mathcal{P}_s(i, T)) \mid \sum_i N(\mathcal{P}_s(i, T)) \leq T \right\}$$

The solution of LP1, denote this FH, is a collection of single arm policies \mathcal{P}_i such that $\text{FH} \leq \sum_i R(\mathcal{P}_i)$ and $\sum_i N(\mathcal{P}_i) \leq T$. We can compute these policies efficiently; see [20, 18, 6, 30].

Rounding to a Feasible Policy: The rounding scheme is now simple. (Variants have appeared before in several stochastic knapsack type optimization contexts [10, 18, 5].) Order the arms in order of $\frac{R(\mathcal{P}_i)}{N(\mathcal{P}_i)}$. Play $\{\mathcal{P}_i\}$ in that order. If the decision in \mathcal{P}_i is to quit, then we move to the next arm. If at any point we have already made T plays, we stop and quit the overall policy.

We first prove a simple algebraic lemma.

Lemma 1. *Suppose r_i/w_i , $i = 1, 2, \dots, n$ is a non-increasing sequence. Let $w = \sum_i w_i$. Then*

$$\sum_{i < k} r_i \left(1 - \frac{1}{w} \sum_{j < i} w_j \right) \geq \frac{(\sum_{i < k} r_i)(\sum_{i < k} w_i)}{2w} + \sum_{i < k} \frac{r_i w_i}{2w} + \sum_{i < k} r_i \frac{\sum_{j \geq k} w_j}{w}$$

Proof. The LHS is $\frac{1}{w}$ of the area under the curve defined by the chain of points $(0, r_1), (w_1, r_1), (w_1, r_1 + r_2), \dots, (\sum_{j < i} w_j, \sum_{j \leq i} r_j), (\sum_{j \leq i} w_j, \sum_{j \leq i} r_j), \dots, (\sum_{j < k-1} w_j, \sum_{j < k} r_j)$, and a horizontal line segment broken into two parts $(\sum_{j < k} w_j, \sum_{j < k} r_j), (w, \sum_{j < k} r_j)$ for convenience.

The non-decreasing ordering implies that all the points in the curve (except the last one) are above the line $y(\sum_{j < k} w_j) = x(\sum_{j < k} r_j)$.

Therefore the area defined in the LHS dominates the sum of the areas defined by (i) the triangle defined by $(0, 0), (\sum_{j < k} w_j, \sum_{j < k} r_j), (0, \sum_{j < k} r_j)$, (ii) the $k-1$ small triangles defined by $(\sum_{j < i} w_j, \sum_{j < i} r_j), (\sum_{j < i} w_j, \sum_{j \leq i} r_j)$, and $(\sum_{j \leq i} w_j, \sum_{j \leq i} r_j)$. and (iii) The rectangle of length $\sum_{j \geq k} w_j$ and height $\sum_{j < k} r_j$.

These are the three terms on the RHS and thus the lemma follows. \square

Theorem 2. *In time $\text{poly}(T, \sum_i |\mathcal{S}_i|)$, we can compute a 2-approximation to the finite horizon Bayesian MAB problem with instantaneous feedback (with budgets and arbitrary priors).*

Proof. We show that the policy described above has an expected reward of $\frac{1}{2}\text{FH}$. Let the number of plays of arm i be T_i . We know $\mathbf{E}[T_i] = N(\mathcal{P}_i)$ and $\sum_i N(\mathcal{P}_i) = T$. We start playing arm i after $\sum_{j < i} T_j$ plays (if the sum is less than T); we apply the Truncation Theorem 1 and the expected reward of \mathcal{P}_i continuing from this point onward is $\left(1 - \frac{1}{T} \min\{T, \sum_{j < i} T_j\}\right) R(\mathcal{P}_i)$. Note that this is a consequence of the independence of arm i from $\sum_{j < i} T_j$. Thus the total expected reward is

$$\begin{aligned} R &= \mathbf{E} \left[\sum_i \left(1 - \frac{1}{T} \min\{T, \sum_{j < i} T_j\} \right) R(\mathcal{P}_i) \right] \geq \mathbf{E} \left[\sum_i \left(1 - \frac{1}{T} \sum_{j < i} T_j \right) R(\mathcal{P}_i) \right] \\ &= \sum_i \left(1 - \frac{1}{T} \sum_{j < i} \mathbf{E}[T_j] \right) R(\mathcal{P}_i) = \sum_i R(\mathcal{P}_i) \left(1 - \frac{\sum_{j < i} N(\mathcal{P}_j)}{T} \right) \end{aligned}$$

By Lemma 1 (with $r_i = R(\mathcal{P}_i)$, $w_i = N(\mathcal{P}_i)$ and $k = n+1$) this is at least $\frac{1}{2} \sum_j R(\mathcal{P}_j) = \frac{1}{2}\text{FH}$. \square

Tight example of the analysis. We show that the gap of the optimum policy and LP1 is a factor of $2 - O(\frac{1}{n})$, even for unit length plays. Consider the following situation: We have two “types” of arms. The type I arm gives a reward 0 with probability $a = 1/n$ and 1 otherwise. The type II arm always gives a reward 0. We have n independent arms. Each has an identical prior distribution of being type I with probability $p = 1/n^2$ and type II otherwise. Set $T = n$.

Consider the symmetric LP solution that allocates one play to each arm; if it observed a 1, it plays the arm for n steps. The expected number of plays made is $n + O(1/n)$, and the expected reward is $n \times 1/n = 1$. Therefore, $\text{LP1} \geq 1 - O(1/n)$.

Consider the optimum policy. We first observe that if the policy ever sees a reward 1 then the optimum policy has found one of the type II arms, and the policy will continue to play this arm for the rest of the time horizon. At any point of time before the time horizon, since $T = n$, there is always at least one arm which has not been played yet. Suppose we play an arm and observe the reward 0, then the posterior probability of this arm being type II increased. So the optimum policy should not prefer this currently played arm over an unplayed arm. Thus the optimum policy would be to order the arms arbitrarily and make a single play on every new arm. If the outcome is 0, the policy quits, otherwise the policy keeps playing the arm for the rest of the horizon. The reward of the optimum policy can thus be bounded by $\sum_{x=0}^{T-1} ap(1+(T-x-1)a) = pa^2T(T+1)/2 + (1-a)/n = \frac{1}{2} + O(\frac{1}{n})$. Thus the gap is a factor of $2 - O(\frac{1}{n})$.

4 Multi-armed Bandits with Delayed Feedback

Define $Q_i = T/\delta_i$. We assume that $\delta_i = o(T/\log T)$, which implies that the horizon is slightly separated from the delays. We show the following theorem.

Theorem 3. *Assuming $\delta_i = o(T/\log T)$, there is a constant factor approximation to the Bayesian MAB problem with delayed feedback. The running time for computing this policy is $\text{poly}(T, \sum_i |\mathcal{S}_i|)$.*

As in Section 3.1, we will use (LP1) to bound of the reward of the best collection of single-arm policies. However, it is not clear how to solve the above LP in polynomial time, since each $\mathcal{P}_s(i, T)$ has description which is exponential in the delay parameter δ_i . We first simplify the structure of the single-arm policies to enable poly-time computation (Step 1). Even after this, it could happen that the policies make most of their plays after $T/2$ steps, so that the truncation theorem cannot be applied directly. We then show how to compact the policies (Step 2) to reduce the horizon - this compaction uses an interesting simulation argument. We then truncate the policies, and solve the LP relaxation over these well-structured policies (Step 3). We finally design a scheduling algorithm for combining these single-arm policies (Step 4) and analyze the approximation ratio.

4.1 Step 1: Block Structured Policies

Definition 2. *A single-arm policy is said to be Block Structured if the policy executes in phases of size $(2\delta_i + 1)$. At the start of each phase (or block), the policy makes at most $\delta_i + 1$ consecutive plays. The policy then waits for the rest of the block in order to obtain feedback on these plays, and then moves to the next block. A block is defined to be full if exactly $\delta_i + 1$ plays are made in it.*

We first show that all single-arm policies can be replaced with block structured policies while violating the time horizon by a constant factor. The idea behind this proof is simple – we simply insert delays of length δ_i after every chunk of plays of length δ_i .

Lemma 2. *Any policy $\mathcal{P}(i, T)$ can be converted it to a Block Structured policy $\mathcal{P}'(i, 2T)$ such that $R(\mathcal{P}(i, T)) \leq R(\mathcal{P}'(i, 2T))$ and $N(\mathcal{P}'(i, 2T)) \leq N(\mathcal{P}(i, T))$*

Proof. We can assume that the policy makes a play at the very first time step, because we can eliminate any wait without any change of behavior of the policy.

Consider the actions of $\mathcal{P}(i, T)$ for the first $\delta_i + 1$ steps, the result of any play in these steps is not known before all these plays are made. An equivalent policy \mathcal{P}' simulates \mathcal{P} for the first $\delta_i + 1$ steps, and then waits for δ_i steps, for a total of $2\delta_i + 1$ steps. This ensures that \mathcal{P}' knows the outcome of the plays before the next block begins.

Now consider the steps from $\delta_i + 2$ to $2\delta_i + 3$ of $\mathcal{P}(i, T)$. As \mathcal{P} is executed, it makes some plays possibly in an adaptive fashion based on the outcome of the plays in the previous $\delta_i + 1$ steps, but not on the current $\delta_i + 1$ steps. \mathcal{P}' however knows the outcome of the previous plays, and can simulate \mathcal{P} for these $\delta_i + 1$ steps and then wait for δ_i steps again. It is immediate to observe that \mathcal{P}' can simulate \mathcal{P} at the cost of increasing the horizon by a factor of $\frac{2\delta_i+1}{\delta_i+1} < 2$. Observe that in each block of $2\delta_i + 1$, \mathcal{P}' can make all the plays consecutively at the start of the block without any change in behavior. The budgets are also respected in this process. This proves the lemma. \square

4.2 Step 2: Well-Structured Policies and Simulation

The block-structured policies constructed above still suffer from the drawback that too many plays can be made close to the horizon T . Initially, the policy can be conservative and play very few times in each block waiting for feedback. In this part, we show how to compact such policies while preserving the reward and making sure the expected number of plays only increases by a constant factor. Our technique uses the idea of *simulation* - we make more plays initially, but hold on to the outcomes of the extra plays. When the original policy makes plays in a subsequent block, we eliminate these plays and instead use the outcomes of the saved up plays. As with the Truncation theorem, this argument crucially uses the martingale property that the reward of the arm are *i.i.d.* draws from the same unknown underlying distribution regardless of when the plays are made.

Definition 3. Define a block-structured policy to be c -delay-free for $c \leq 1$ if the first time the policy encounters a block with at least $c\delta_i$ plays, it plays every step (without waiting) beyond this point (using feedback from δ_i plays ago) until it stops executing.

Intuitively, we are eliminating delays if a policy makes a sufficiently large number of plays in a block. This step not only compacts the policy, but also shrinks the state space significantly, since if the policy executes in a delay-free fashion, we can pretend the feedback is instantaneous. (In reality, the feedback is from δ_i plays ago, but we can easily couple the two executions.) We now show that any block-structured policy can be made c -delay-free.

Lemma 3. Given any Block Structured policy $\mathcal{P}(i, 2T)$ we can construct a c -delay-free Block-structured policy $\mathcal{P}'(i, 2T)$, such that $R(\mathcal{P}(i, 2T)) \leq R(\mathcal{P}'(i, 2T))$ and $N(\mathcal{P}'(i, 2T)) \leq (1 + \frac{1}{c})N(\mathcal{P}(i, 2T))$.

Proof. Consider the first time the policy encounters a block with $c\delta_i$ plays on some decision path. We play the arm continuously beyond this point and *simulate* the behavior of the original policy \mathcal{P} . For each contiguous play the policy \mathcal{P}' makes, this outcome is available after δ_i steps. Simulate the next play of \mathcal{P} using this outcome, until \mathcal{P} stops execution. Clearly, \mathcal{P}' makes at most δ_i additional plays than \mathcal{P} on any decision path; since \mathcal{P} made at least $c\delta_i$ plays, this shows that $N(\mathcal{P}'(i, 2T)) \leq (1 + \frac{1}{c})N(\mathcal{P}(i, 2T))$. Since the execution of \mathcal{P}' is coupled play-by-play with the execution of \mathcal{P} , it is clear that $R(\mathcal{P}(i, 2T)) \leq R(\mathcal{P}'(i, 2T))$. \square

The policies constructed above still suffer from being too sparse if the number of plays in each block is less than $c\delta_i$. In this step, we will compact these policies further to retain mostly the full blocks. Note that any policy uses at most $Q_i = \frac{T}{\delta_i} = \omega(\log \delta_i)$ blocks.

Definition 4. For constant $\alpha < 1$, define a c -delay-free block-structured single-arm policy \mathcal{P} to be (α, c) -well-structured if after encountering at most $q = (\alpha + o(1))Q_i$ blocks, the policy switches to playing continuously, i.e., executing in delay-free mode.

The next lemma further compacts the block-structured policies into well-structured policies.

Lemma 4. *For any $\alpha < 1$ and $c \leq \frac{\alpha}{\alpha+2}$, given a c -delay-free policy $\mathcal{P}(i, 2T)$, there is a (α, c) -well-structured policy \mathcal{P}' such that $R(\mathcal{P}(i, 2T)) \leq R(\mathcal{P}'(i, 2T))$ and $N(\mathcal{P}'(i, 2T)) \leq (1 + \frac{2}{\alpha})N(\mathcal{P}(i, 2T))$.*

Proof. Consider the execution of \mathcal{P} . All blocks in this policy have at most $c\delta_i$ plays, unless the policy is executing in delay-free mode. We consider the execution of the original policy, and show a coupled execution in the new policy \mathcal{P}' so that if on a decision path, \mathcal{P} used k blocks, then the number of blocks on the same decision path in \mathcal{P}' is $\alpha k + O(\log \delta_i)$. We group blocks into size classes; size class s has blocks whose number of plays lies in $[2^s, 2^{s+1}]$.

We couple the executions as follows: Consider the decision tree \mathcal{P} top-down. Increase the plays in the root block r by a factor of $(1 + \frac{2}{\alpha})$; suppose there were x plays originally in this block, and the size class of this block is s . Consider the first $\frac{1}{\alpha}$ blocks on each decision path downstream of r whose size class is at most s ; eliminate these blocks and use the outcomes of the $\frac{2x}{\alpha}$ extra plays made at r to simulate the behavior of \mathcal{P} in these blocks as follows. Simply store the outcomes of the extra plays at r (without updating the prior), and use these outcomes when \mathcal{P} makes those plays. Since the underlying reward distribution is fixed, these plays will be stochastically identical. We note that \mathcal{P}' does not really need to “store” the outcomes - it can simply update the prior at r and follow the execution of \mathcal{P} as if it had not updated it. Again, since the reward process is a martingale, both these executions are stochastically identical.

On each decision path, either we used r to eliminate $\frac{1}{\alpha}$ blocks, or there are no blocks left with size class at most s . Now repeat this procedure on the roots of each decision sub-tree downstream of r . When this procedure terminates, consider any decision path, and mark the leftover blocks whose increase in plays could not eliminate exactly $\frac{1}{\alpha}$ blocks. For any size class s , it is clear that there can be at one such block, else the earlier block could have eliminated the later one. Therefore, if a decision path at k blocks in \mathcal{P} , the above procedure shrinks the number of blocks to at most $\alpha k + \frac{\log_2 \delta_i}{\alpha}$, since there are at most $\log_2 \delta_i$ size classes. This procedure increases the expected number of plays by at most a factor of $1 + \frac{2}{\alpha}$, and we have a guarantee that each decision path makes at most $(\alpha + o(1))Q_i$ plays, since $\log \delta_i = o(Q_i)$. \square

4.3 Step 3: Truncation and Solving the LP Relaxation

We now truncate the horizon of the well-structured policies losing a constant factor in the reward.

Lemma 5. *Given any $(1/8, 1/17)$ -well-structured single arm policy $\mathcal{P} = \mathcal{P}(i, 2T)$, there is an identical policy that stops execution after $\frac{T}{2}$ times steps (denoted by \mathcal{P}'), that satisfies the following: (i) $R(\mathcal{P}') = \frac{R(\mathcal{P})}{8}$ and (ii) $N(\mathcal{P}') \leq N(\mathcal{P})$.*

Proof. On any decision path in \mathcal{P} , there can be at most $Q_i(1/8 + o(1))$ blocks before the policy executes in delay-free mode. These blocks take up at most $T/4$ time steps, and the plays beyond that are delay-free. This implies that if the policy’s horizon is truncated to $T/2$ time steps, and the policy has not stopped executing, it makes at least $T/4$ contiguous plays, whereas \mathcal{P} could have made at most $2T$ plays. Therefore, on any decision path, the number of plays is reduced by at most a factor of 8, and by the Truncation theorem 1, the lemma follows. \square

We term well-structured policies that terminate on encountering more than $T/2$ time steps as **truncated well-structured** policies. The above lemmas can be summarized as:

Theorem 4. *(LP1) has a $1/\alpha$ -approximation over truncated $(\alpha, \frac{\alpha}{\alpha+2})$ -well structured policies for $\alpha \leq 1/8$. This LP has the relaxed constraint: $\sum_i N(\mathcal{P}_s(i, T/2)) \leq \gamma T$ where $\gamma = 2(1+1/\alpha)(1+2/\alpha)$.*

We now formulate a new LP relaxation over truncated well-structured policies constructed in Theorem 4. This new LP is polynomial over $|\mathcal{S}_i|$ and we avoid the exponential dependence on δ_i due to the above transformations. Towards this end, we explicitly write out the state space of a truncated (α, c) -well-structured policy for a single arm. For $u \in \mathcal{S}_i$, we create a state $W(u, t)$ for $1 \leq t \leq T/2$. The meaning of this state is as follows: The posterior at the beginning of a block is $u \in \mathcal{S}_i$, and t time steps have elapsed so far. This block is either regular (in which case, at most αQ_i blocks have elapsed so far), or no-delay. For no-delay blocks, we can assume the block-length is 1 and feedback is instantaneously available; this unifies the presentation of the LP. (The feedback is actually from plays made δ_i steps ago, but we can couple the two executions.)

For any state $\sigma = W(u, t)$ of arm i , define the following quantities. These quantities are an easy computation from the description of \mathcal{S}_i , and the details are omitted.

1. Let $r_i(\sigma, \ell)$ denote the expected reward obtained when consecutive ℓ plays are made at state σ , and feedback is obtained at the very end. Note that for regular state σ , we have $0 \leq \ell c \delta_i$. For a no-delay state, $0 \leq \ell \leq 1$.
2. Let $p_i(\sigma, \sigma', \ell)$ denote the probability that if the state at the beginning of a block is σ , and ℓ plays are made in the block, the state at the beginning of the next block is σ' .

We will formulate an LP to find a randomized truncated well-structured policy \mathcal{P} . Define the following variables over the decision tree of the policy:

- $x_{i\sigma}$: the probability that the state for arm i at the start of a block is σ .
- $y_{i\sigma\ell}$: probability that the policy for arm i makes ℓ consecutive plays starting at state σ .

We have the following LP relaxation, which simply encodes finding one randomized well-structured policy per arm so that the expected number of plays made is at most γT . The objective to this LP is $\Omega(OPT)$ as a consequence of Theorem 4.

$$\begin{aligned}
LP2 = \text{Maximize } & \sum_{i, \sigma, \ell} r_i(\sigma, \ell) y_{i\sigma\ell} \\
& \sum_{\sigma} x_{i\sigma} \leq 1 \quad \forall i \\
& \sum_{\ell} y_{i\sigma\ell} \leq x_{i\sigma} \quad \forall i, \sigma \\
& \sum_{\sigma, \ell} y_{i\sigma\ell} p_i(\sigma, \sigma', \ell) = x_{i\sigma'} \quad \forall i, \sigma' \\
& \sum_{i, \sigma, \ell} \ell y_{i\sigma\ell} \leq \gamma T \\
& x_{i\sigma}, y_{i\sigma\ell} \geq 0 \quad \forall i, \sigma, \ell
\end{aligned}$$

The LP policy: We scale down the LP solution by a factor of γ so that the final constraint becomes $\sum_{i, \sigma, \ell} \ell y_{i\sigma\ell} \leq T$; denote the new LP as $(LP2s)$ for convenience. $(LP2s)$ yields one randomized well-structured policy $\mathcal{P}^r(i, T/2)$ for each arm i . This policy is succinctly described in Figure 1.

Policy $\mathcal{P}^r(i, T/2)$:

- If the state at the beginning of a block is σ :
 1. Choose n with probability $\frac{y_{i\sigma n}}{x_{i\sigma}}$, and make n plays in the current block.
 2. Wait till the end of the block; obtain feedback for the n plays; and update state.

Figure 1: Single-arm policy $\mathcal{P}^r(i, T/2)$.

4.4 Step 4: Priority Based Combining of Different Policies

At this point we have a collection of randomized policies $\mathcal{P}^r(i, T/2)$ such that $\sum_i N(\mathcal{P}^r(i, T/2)) \leq T$ and $\sum_i R(\mathcal{P}^r(i, T/2)) = \Omega(OPT)$. We now show how to combine the single arm policies $\mathcal{P}^r(i, T/2)$ to achieve a globally approximate and feasible solution. This step is not complicated but we note that multiple policies must remain active in the combination. This aspect necessitates a novel priority based scheme for combining the policies.

Consider the execution of $\mathcal{P}^r(i, T/2)$. We describe the arm as *active* if it is either making plays or is at the beginning of a block where it can make plays; and *passive* if it is waiting for feedback on the plays within the block. Any arm which completed its waiting for the feedback turns from passive to active mode. The final policy is shown in Figure 2.

1. Choose an arbitrary order the arms $\{i\}$ denoted by π .
2. Each arm “participates” with probability $1/4$. Initially, all participating arms are active.
3. On each new play, among all participating arms:
 - (a) Find the lowest rank arm, say i' , that is active.
 - (b) Allocate the current play to i' according to the policy $\mathcal{P}^r(i', T/2)$.
 - (c) (As a result of this allocation, the arm may become passive and wait for feedback.)

Figure 2: The final policy **Combine**.

The next lemma follows from an application of Markov’s inequality in the same spirit as prior work without delayed feedback [18].

Lemma 6. *The expected contribution of i from the combined policy is at least $R(\mathcal{P}^r(i, T/2))/8$.*

Proof. First observe that i being active does not impact any play of G_i . The expected number of plays made by G_i is $\sum_{i' \in G_i} N(\mathcal{P}^r(i', T/2))/4 \leq T/4$. Thus by Markov’s Inequality, the arms in G_i make $T/2$ plays with probability at most $1/2$. Next note that with probability $1/8$, arm i is active and in a scenario where the number of plays made by bidders before i is at most $T/2$. This means **every** decision path of $\mathcal{P}^r(i, T/2)$ can complete in this scenario (since i gets higher priority than all bidders following it). Thus with probability $1/8$ the policy $\mathcal{P}^r(i, T/2)$ executes without any interference. Thus the expected contribution from arm i is at least $R(\mathcal{P}^r(i, T/2))/8$. \square

The above lemma shows Theorem 3 by linearity of expectation, concluding the analysis.

References

- [1] D. Agarwal, B.-C. Chen, and P. Elango. Explore/exploit schemes for web content optimization. *Proceedings of ICDM*, pages 1–10, 2009.
- [2] T. W. Anderson. Sequential analysis with delayed observations. *Journal of the American Statistical Association*, 59(308):1006–1015, 1964.
- [3] P. Armitage. The search for optimality in clinical trials. *International Statistical Review*, 53(1):15–24, 1985.
- [4] P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2-3):235–256, 2002.
- [5] N. Bansal, A. Gupta, J. Li, J. Mestre, V. Nagarajan, and A. Rudra. When lp is the cure for your matching woes: Improved bounds for stochastic matchings. *CoRR*, abs/1008.5356, 2010.
- [6] D. Bertsekas. *Dynamic Programming and Optimal Control*. Athena Scientific, second edition, 2001.
- [7] N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning and Games*. Cambridge University Press, 2006.
- [8] D. Chakrabarty and G. Goel. On the approximability of budgeted allocations and improved lower bounds for submodular welfare maximization and gap. *Proceedings of FOCS*, pages 687–696, 2008.
- [9] S. C. Choi and V. A. Clark. Sequential decision for a binomial parameter with delayed observations. *Biometrics*, 26(3):411–420, 1970.
- [10] B. C. Dean, M. X. Goemans, and J. Vondrak. Approximating the stochastic knapsack problem: The benefit of adaptivity. *Math. of Operations Research*, 33(4):945–964, 2008.
- [11] N. R. Devanur and T. P. Hayes. The adwords problem: Online keyword matching with budgeted bidders under random permutations. In *Proc. 10th Annual ACM Conference on Electronic Commerce (EC)*, 2009.
- [12] S. Ehrenfeld. On a scheduling problem in sequential analysis. *The Annals of Mathematical Statistics*, 41(4):1206–1216, 1970.
- [13] S. G. Eick. The two-armed bandit with delayed responses. *The Annals of Statistics*, 16(1):254–264, 1988.
- [14] V. F. Farias and R. Madan. Irrevocable multi-armed bandit policies. *Operations Research (to appear)*, available at http://web.mit.edu/vivekf/www/papers/MAB_final.pdf.
- [15] J. Feldman, N. Korula, V. S. Mirrokni, S. Muthukrishnan, and M. Pl. Online ad assignment with free disposal. *Proc. of WINE*, pages 374–385, 2009.
- [16] J. C. Gittins and D. M. Jones. A dynamic allocation index for the sequential design of experiments. *Progress in statistics (European Meeting of Statisticians)*, 1972.
- [17] A. Goel, S. Khanna, and B. Null. The ratio index for budgeted learning, with applications. *Proceedings of SODA*, 2009.

- [18] S. Guha and K. Munagala. Multi-armed bandits with metric switching costs. *Proceedings of ICALP*, 2009.
- [19] S. Guha and K. Munagala. MAXMAB: Budgeted multi-armed bandits with concave objectives. *Manuscript*, 2011.
- [20] S. Guha, K. Munagala, and P. Shi. Approximation algorithms for restless bandit problems. *CoRR*, abs/0711.3861, 2007.
- [21] T. L. Lai and H. Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6:4–22, 1985.
- [22] O. Madani, D. J. Lizotte, and R. Greiner. Active model selection. In *UAI '04: Proc. 20th Conf. on Uncertainty in Artificial Intelligence*, pages 357–365, 2004.
- [23] A. Mehta, A. Saberi, U. V. Vazirani, and V. V. Vazirani. Adwords and generalized on-line matching. *J. ACM*, 54(5), 2007.
- [24] J. L. Ny, M. Dahleh, and E. Feron. Multi-UAV dynamic routing with partial observations using restless bandits allocation indices. In *Proceedings of the 2008 American Control Conference*, 2008.
- [25] S. Pandey and C. Olston. Handling advertisements of unknown quality in search advertising. *Proceedings of NIPS*, pages 1065–1072, 2006.
- [26] H. Robbins. Some aspects of the sequential design of experiments. *Bulletin American Mathematical Society*, 55:527–535, 1952.
- [27] J. Schneider and A. Moore. Active learning in discrete input spaces. In *34th Interface Symp.*, 2002.
- [28] R. Simon. Adaptive treatment assignment methods and clinical trials. *Biometrics*, 33(4):743–749, 1977.
- [29] Y. Suzuki. On sequential decision problems with delayed observations. *Annals of the Institute of Statistical Mathematics*, 18(1):229–267, 1966.
- [30] J. N. Tsitsiklis. A short proof of the Gittins index theorem. *Annals of Appl. Prob.*, 4(1):194–199, 1994.
- [31] A. Wald. *Sequential Analysis*. Wiley, New York, 1947.

A Budgeted Allocation with Delayed Feedback

Recall that the budgeted allocation problem is a collection of MAB instances, one for each impression type. For each impression of type j , the MAB instance $M(T_j)$ is over T_j arrivals (or plays) of this impression. (Note that $\sum_j T_j = T$.) The arms for this impression are denoted (i, j) , where i denotes the bidders. If a play along (i, j) is successful (which is observed after δ_{ij} arrivals of impressions of type j), bidder i accrues reward b_{ij} ; the probability of success is drawn from an independent prior \mathcal{D}_{ij} . The arms are connected by the budget constraint stating that the total reward accrued by bidder i across all MAB instances is at most B_i .

Algorithm and Analysis: The overall algorithm is similar to that for a single MAB instance.

1. The initial LP relaxation (that we don't solve) encodes constructing one randomized policy per arm (i, j) such that: (i) The total expected number of plays for MAB M_j is at most T_j ; and (ii) The total expected reward accrued for bidder i is at most B_i . Note that the latter constraint was not present in (LP1) in the previous section.
2. With a loss of factor 2 in the reward, we define "shadow budgets" $B_{ij} \leq B_i$ such that B_{ij}/b_{ij} are integers, as discussed in Section 3.
3. For each (i, j) , we transform the single-arm policy (with the budgets B_{ij}) just as in the previous section into a well-structured policy using $T_j/2$ impressions.
4. We solve an LP over truncated well-structured policies, with the additional constraint that the total reward accrued for bidder i is at most B_i . Let the LP objective value be M .
5. To construct the global policy, we run the policy **Combine** independently for each j . Let the randomized well-structured policy for arm (i, j) that is constructed by the LP be denoted $\mathcal{P}^r(i, j, T_j/2)$, and let Z_{ij} denote the reward achieved. Note $Z_{ij} \leq B_{ij} \leq B_i$.

The analysis in Section 3 shows that the policy **Combine** yields expected reward $\sum_{i,j} \mathbf{E}[Z_{ij}] = \Omega(M)$. However we can only extract a reward of $\min\{B_i, \sum_j Z_{ij}\}$ from advertiser i . We show:

Lemma 7. *For a collection of random variables $\{Z_{ij}\}$, if $\sum_j \mathbf{E}[Z_{ij}] \leq B_i$ and $Z_{ij} \leq B_i$ then $\mathbf{E}[\min\{B_i, \sum_j Z_{ij}\}] \geq \frac{1}{2} \sum_j \mathbf{E}[Z_{ij}]$.*

Proof. Consider a process where advertiser i settles the bill for j , before proceeding to the bill for $j+1$. If the payout on j is Y_j then $Y_j = \min\left\{Z_{ij}, \max\{B_i - \sum_{j' < j} Z_{ij'}, 0\}\right\}$.

$$Y_j = B_i \min\left\{\frac{Z_{ij}}{B_i}, \max\left(1 - \frac{\sum_{j' < j} Z_{ij'}}{B_i}, 0\right)\right\} \geq \max\left(1 - \frac{\sum_{j' < j} Z_{ij'}}{B_i}, 0\right) Z_{ij}$$

The last part follows from $\min\{a, b\} \geq ab$, for any $0 \leq a, b \leq 1$. Clearly $\sum_j Y_j \leq B_i$ in every scenario. And as a consequence of Y_j being independent of $Z_{ij'}$ for $j' < j$:

$$\mathbf{E}[Y_j] \geq \mathbf{E}\left[\max\left(1 - \frac{\sum_{j' < j} Z_{ij'}}{B_i}, 0\right)\right] \mathbf{E}[Z_{ij}] \geq \left(1 - \frac{\sum_{j' < j} \mathbf{E}[Z_{ij'}]}{B_i}\right) \mathbf{E}[Z_{ij}]$$

We can estimate $\sum_j \mathbf{E}[Y_j]$ to be at least $\frac{1}{B_i}$ times the area under the staircase defined by

$$(0, \mathbf{E}[Z_{i1}]), \dots, \left(\sum_{j' < j} \mathbf{E}[Z_{ij'}], \sum_{j' \leq j} \mathbf{E}[Z_{ij'}]\right), \left(\sum_{j' \leq j} \mathbf{E}[Z_{ij'}], \sum_{j' \leq j} \mathbf{E}[Z_{ij'}]\right), \dots, \left(\sum_j \mathbf{E}[Z_{ij}], \sum_j \mathbf{E}[Z_{ij}]\right), (B_i, \sum_j \mathbf{E}[Z_{ij}])$$

This area includes the triangle defined $(0, 0)$, $(B_i, 0)$ and $(B_i, \sum_j \mathbf{E}[Z_{ij}])$ and thus the reward is therefore at least $\frac{1}{B_i} \frac{1}{2} B_i \sum_j \mathbf{E}[Z_{ij}] = \frac{1}{2} \sum_j \mathbf{E}[Z_{ij}]$. This proves the lemma. \square

As a consequence of Lemma 7, and that $M = \Omega(OPT)$ we have:

Theorem 5. *If $T_j \geq 48(8\delta_{ij} + 2) \log T_j$ for all i, j then in time polynomial in n, m , and $\max_{i,j} |\mathcal{S}_{ij}|$, we can compute a policy whose expected reward is a $O(1)$ approximation to the expected reward of the optimal policy for the Bayesian budgeted allocation problem with delayed feedback.*

B Example of Priors: Beta Distributions

Consider a fixed but unknown distribution over two outcomes (success or failure of treatment, click or no click, conversion or no conversion). This is a Bernoulli($1, \theta$) trial where θ is unknown. One way of encoding the uncertainty about θ is to use a distribution which encodes our “prior” belief regarding θ . However on getting a new sample, we update our belief to a “posterior distribution” using Bayes’ rule. Note that we need a family of distributions and associated update rules to specify the prior. For two valued outcomes, one such family is the Beta distribution, referred to as the conjugate prior of the Bernoulli distribution. A Beta distribution with parameters $\alpha_0, \alpha_1 \in \mathbb{Z}^+$, which we denote $Beta(\alpha_1, \alpha_0)$ has p.d.f. of the form $c\theta^{\alpha_1-1}(1-\theta)^{\alpha_0-1}$, where c is a normalizing constant. $Beta(1, 1)$ is the uniform distribution. The distribution $Beta(\alpha_1, \alpha_0)$ corresponds to the current (posterior) distribution over the possible values of θ after having observed $(\alpha_1 - 1)$ 1’s and $(\alpha_0 - 1)$ 0’s, starting from the belief that θ was uniform, distributed as $Beta(1, 1)$. Given the distribution $Beta(\alpha_1, \alpha_0)$ as our prior, the expected value of θ is $\frac{\alpha_1}{\alpha_1 + \alpha_0}$. In this example, updating the prior on a sample is straightforward. On seeing a 1, the posterior (of the current sample, and the prior for the next sample) is $Beta(\alpha_1 + 1, \alpha_0)$. On seeing a 0, the new distribution is $Beta(\alpha_1, \alpha_0 + 1)$. As stated earlier, the hard case for our algorithms (and the case that is typical in practice) is when the input to the problem is a set of arms $\{i\}$ with priors $\mathcal{D}_i \sim Beta(\alpha_{1i}, \alpha_{0i})$ where $\alpha_{0i} \gg \alpha_{1i}$ which corresponds to a set of **poor prior expectations** of the arms.

Representing Priors as a DAG. Given a family of priors we model them as a directed acyclic graph (DAG) where the vertices correspond to the priors and posteriors. The root ρ encodes the initial prior \mathcal{D} . The children of ρ correspond to the possible observations of a sample drawn from \mathcal{D} . In the case of Beta priors and two valued outcomes, a node u corresponding to $Beta(\alpha_1, \alpha_0)$ would have two children corresponding to $Beta(\alpha_1 + 1, \alpha_0)$, $Beta(\alpha_1, \alpha_0 + 1)$ and the edges will be labeled 1, 0 respectively. The probability of taking the edge labeled 1 from u to the node v corresponding to $Beta(\alpha_1 + 1, \alpha_0)$ would be the probability of observing a 1 given the prior $Beta(\alpha_1, \alpha_0)$, which is $\frac{\alpha_1}{\alpha_1 + \alpha_0}$. We will define $\mathbf{p}_{uv} = \frac{\alpha_1}{\alpha_1 + \alpha_0}$. At each node u the probability of observing 1 corresponds to the “reward” μ_u . Thus the DAG represents the evolution of priors/posteriors completely, and further (for two valued outcomes) each node is represented by a pair of numbers. Therefore for a horizon of size T , the entire information is captured by a DAG of size $O(T^2)$.

Martingale Property: A *critical* observation, true of all conjugate prior distributions and the resulting DAGs, is that we have a **Martingale property**: $\mu_u = \sum_{v \in \mathcal{S}_i} \mu_v \mathbf{p}_{uv}$ (summing over the children of u). *Therefore this gives us a way of switching between two different views of the same evolution of the posterior space:* the first view is where the p_i (equivalently D_i) is chosen according to \mathcal{D}_i and we observe a collection of trajectories and take expectation over them, and the second view is where we have a Martingale process of the posterior spaces in the DAG.